Une suite logicielle pour la protéomique interfacée sur une grille de calcul. Utilisation d'algorithmes libres pour l'identification MS/MS, le séquençage de novo et l'annotation fonctionnelle.

Christine Carapito, Jérôme Pansanel, Patrick Guterl, Alexandre Burel, Fabrice Bertile, Stéphane Genaud, Alain van Dorsselaer, Christelle Roy

Une suite logicielle pour la protéomique interfacée sur une grille de calcul. Utilisation d'algorithmes libres pour l'identification MS/MS, le séquençage *de novo* et l'annotation fonctionnelle.

Auteurs
Christine CARAPITO[1], Jérôme PANSANEL[2], Patrick GUTERL[1], Alexandre BUREL[1], Fabrice BERTILE[1], Stéphane GENAUD[3], Alain VAN DORSSELAER[1], Christelle ROY[2]

Affiliations
[1] Laboratoire de Spectrométrie de Masse BioOrganique, DSA, IPHC, UMR7178, CNRS, Université de Strasbourg, 25 rue Becquerel 67087 Strasbourg, France
[2] Département Recherches Subatomiques, DRS, IPHC, UMR7178, CNRS, Université de Strasbourg, 23 rue du Loess 67037 Strasbourg, France
[3] Laboratoire de Sciences de l'Image, de l'Informatique et de la Télédétection, équipe ICPS (LSIIT / ICPS), Université de Strasbourg, CNRS UMR 7005, Strasbourg, France

Contexte
Les progrès instrumentaux en spectrométrie de masse (MS) de ces 20 dernières années ont conduit au développement d'instruments générant des données MS/MS de plus en plus volumineuses (du fait d'une grande rapidité d'acquisition des spectres de fragmentation).
Par ailleurs, la soumission des résultats d'identification de protéines à partir de ces données MS/MS est de plus en plus réglementée par les journaux du domaine qui recommandent l'utilisation d'algorithmes transparents (open-source) et multiples si possible.
Dans ce contexte, afin de répondre au besoin croissant de puissance de calcul nécessaire à l'interprétation des données MS/MS, une suite logicielle bâtie sur des logiciels libres a été adaptée et améliorée sur une grille de calcul.

Méthodes et Résultats
La suite logicielle développée à l'IPHC permet :
- de créer, d'extraire, de concaténer, de formater des banques de séquences protéiques, notamment à partir des banques de séquences publiques accessibles telles que NCBInr, UniProtKB, UniProtKB/SwissProt, …
- de lancer des requêtes OMSSA (Open Mass Spectrometry Search Algorithm) pour l'identification de protéines à partir de données MS/MS sur la grille de calcul locale de l'IPHC (1024 cœurs de calculs, site TIER-2 de la grille LHC (Large Hadron Collider)) et mondiale.
- d'interpréter à haut débit des données MS/MS acquises sur des organismes non séquencés par séquençage *de novo* suivi de recherches d'homologies de séquences.
- d'extraire de manière automatisée les annotations fonctionnelles disponibles sur les protéines identifiées.
L'ensemble de ces outils est disponible sur le site **https://msda.u-strasbg.fr.**

Conclusion
L'adaptation de la suite logicielle sur une grille de calcul permet de répondre aux importants besoins de puissance de calcul non accessibles à ce jour dans les laboratoires de protéomique. En effet, selon le type de requête (nombre de spectres MS/MS, taille de la banque de séquences, recherche de modifications post-traductionnelles, séquençage *de novo*, …), un gain d'un facteur 100, voire 1000 est obtenu en routine grâce à la parallélisation des outils et à leur lancement sur une grille de calcul.